PCP
PLANT & CELL PHYSIOLOGY

Triticeae Resources in Ensembl Plants

Dan M. Bolser, Arnaud Kerhornou, Brandon Walts and Paul Kersey*

Ensembl Genomes, EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK
*Corresponding author: E-mail, pkersey@ebi.ac.uk; Fax, +44 223 494 468.
(Received September 22, 2014; Accepted November 12, 2014)

Recent developments in DNA sequencing have enabled the large and complex genomes of many crop species to be determined for the first time, even those previously intractable due to their polyploid nature. Indeed, over the course of the last 2 years, the genome sequences of several commercially important cereals, notably barley and bread wheat, have become available, as well as those of related wild species. While still incomplete, comparison with other, more completely assembled species suggests that coverage of genic regions is likely to be high. Ensembl Plants (http://plants.ensembl.org) is an integrative resource organizing, analyzing and visualizing genome-scale information for important crop and model plants. Available data include reference genome sequence, variant loci, gene models and functional annotation. For variant loci, individual and population genotypes, linkage information and, where available, phenotypic information are shown. Comparative analyses are performed on DNA and protein sequence alignments. The resulting genome alignments and gene trees, representing the implied evolutionary history of the gene family, are made available for visualization and analysis. Driven by the case of bread wheat, specific extensions to the analysis pipelines and web interface have recently been developed to support polyploid genomes. Data in Ensembl Plants is accessible through a genome browser incorporating various specialist interfaces for different data types, and through a variety of additional methods for programmatic access and data mining. These interfaces are consistent with those offered through the Ensembl interface for the genomes of non-plant species, including those of plant pathogens, pests and pollinators, facilitating the study of the plant in its environment.

Keywords: Comparative genomics • Functional genomics • Genetic variation • Genome browser • Transcriptomics • Triticeae.

Abbreviations: API, Application Programming Interface; EST, expressed sequence tag; FTP, File Transfer Protocol; GO, gene ontology; IEA, Inferred by Electronic Annotation; MIPS, Helmholtz Zentrum München; POPSEQ, POPulation SEQuencing; QTL, quantitative trait locus; REST, REpresentational State Transfer; RNA-Seq, RNA sequencing; SNP, single nucleotide polymorphism; SQL, structured query language; TREP, Triticeae Repeat Sequence Database.

Introduction

The first cereal genome sequenced was rice in 2002 (Goff et al. 2002, Yu et al. 2002). More recently, progress has accelerated with the publication of the genome sequence of maize in 2009 (Schnable et al. 2009), barley in 2012 (International Barley Sequencing Consortium 2012), progenitors of the bread wheat A and D genomes in 2013 (Jia et al. 2013, Ling et al. 2013) and the draft bread wheat genome itself in 2014 (Brenchley et al. 2012, International Wheat Genome Sequencing Consortium 2014). These four cereals, barley, maize, rice and wheat, together account for 30% of global food production or 2.4 out of 3.8 billion tonnes annually.

It is important to note that the current reference genome assemblies vary considerably in their contiguity and in the detail of available functional annotation. The Triticeae genomes were all sequenced primarily using short read sequencing (mainly from the Illumina platform), and the completion of these assemblies remains a scientific challenge, due to their large size and repetitive nature. The improvement of sequencing technologies, particularly those capable of capturing long-range information, will be necessary to achieve this goal. However, even in their existing condition, these resources are already sufficiently complete to be usefully represented through data analysis and visualization platforms designed for genomes with finished assemblies, such as Ensembl Plants.

Ensembl Plants (http://plants.ensembl.org) offers integrative access to a wide range of genome-scale data from plant species (Kersey et al. 2014), using the Ensembl software infrastructure (Flicek et al. 2014). Currently, the site includes data from 38 plant genomes, from algae to flowering plants. Genomes are selected for inclusion in the resource based on the availability of the complete genome sequence, their importance as model organisms (e.g. *Arabidopsis thaliana*, *Brachypodium distachyon*), their importance in agriculture (e.g. potato, tomato, various cereals and Brassicaceae) or because of their interest as evolutionary reference points (e.g the basal angiosperm, *Amborella trichopoda*, the aquatic alga *Chlamydomonas reinhardtii*, the moss *Physcomitrella patens* and the vascular non-seed spikemoss *Selaginella moellendorffii*). In total, the resource contains the genomes of 19 true grasses, *Musa accuminata* (banana), 12 dicots and six other species that provide evolutionary context for the plant lineage.

Plant Cell Physiol. 56(1): e3(1–11) (2015) doi:10.1093/pcp/pcu183, Advance Access publication on 27 November 2014, available FREE online at www.pcp.oxfordjournals.org

All species in the resource have data for genome sequence, annotations of protein-coding and non-coding genes, and gene-centric comparative analysis. Additional data types within the resource include gene expression, sequence polymorphism and whole-genome alignments, which are selectively available for different species. In this sense, Ensembl Plants is similar to comparable, species-, clade- or data type-specific resources such as WheatGenome.info (Lai et al. 2012), HapRice (Yonemaru et al. 2014) or ATTED-II (Obayashi et al. 2014).

Ensembl Plants is released 4–5 times a year, in synchrony with releases of other genomes (from animals, fungi, protists and bacteria) in the Ensembl system. The provision of common interfaces allows access to genomic data from across the tree of life in a consistent manner, including data from plant pathogens, pests and pollinators.

Database

The Ensembl genome browser

Interactive access to Ensembl Plants is provided through an advanced genome browser. The browser allows users to visualize a graphical representation of a completely assembled chromosome sequence or a contiguous sequence assembly comprising only a small portion of a molecule at various levels of resolution. Functionally interesting 'features' are depicted on the sequence with defined locations. Features include conceptual annotations such as genes and variant loci, sequence patterns such as repeats, and experimental data such as sequence features mapped onto the genome, which often provide direct support for the annotations (**Fig. 1**). Functional information is provided through import of manual annotation from the UniProt Knowledgebase (Uniprot Consortium 2014), imputation from protein sequence using the classification tool InterProScan (Jones et al. 2014), or by projection from orthologs (described below). Users can download much of the data available on each page in a variety of formats, and tools exist for upload of various types of user data, allowing users to see their own annotation in the context of the reference sequence. DNA- and protein-based sequence search are also available.

All genomes included in Ensembl Plants are periodically run through the Ensembl comparative analysis pipelines, generating DNA and protein sequence alignments. Gene trees are based on protein sequence alignments and show the inferred evolutionary history of each gene family (Vilella et al. 2009). Specialized views are available for these data (e.g. see **Figs. 2**, **5** and **6**), and also for data types including variation (**Fig. 3**), regulation and expression.

The data are stored in MySQL databases using the same schemas as those used by other Ensembl sites. Direct access to these is provided through a public MySQL server and additionally through well-developed Perl and REST APIs. Database dumps and common data sets, such as DNA, RNA, protein sequence sets and sequence alignments, can be directly downloaded in bulk via FTP (ftp://ftp.ensemblgenomes.org).

In addition to the primary databases, Ensembl Plants also provides access to denormalized data warehouses, constructed using the BioMart toolkit (Kasprzyk 2011). These are specialized databases optimized to support the efficient performance of common gene- and variant-centric queries, and can be accessed through their own web-based and programmatic interfaces.

Triticeae genomes in Ensembl Plants

Four Triticeae genomes are currently hosted in Ensembl Plants (**Table 1**): *Hordeum vulgare* (barley), *Triticum aestivum* (bread wheat, also known as common wheat) and the genomes of two of bread wheat's diploid progenitors: *Triticum urartu* (the A-genome progenitor) and *Aegilops tauschii* (the D-genome progenitor). In addition, a further three wheat transcriptomes were included by alignment (described below).

Barley is the world's fourth most important cereal crop and an important model for ecological adaptation, having been cultivated in all temperate regions from the Arctic Circle to the tropics. It was one of the first domesticated cereal grains, originating in the Fertile Crescent of south-west Asia/north-east Africa >10,000 years ago (Harlan and Zohary 1966). With a haploid genome size of approximately 5.3 Gbp in seven chromosomes, the barley genome is among the largest yet sequenced. However, as a diploid, it is a natural model for the genetics and genomics of the polyploid members of the Triticeae tribe, including wheat and rye.

The current barley genome assembly (cv. Morex) was produced by the International Barley Genome Sequencing Consortium (2012). The assembly is highly fragmented, but comparison with related grass species suggested that coverage of the gene space was good. The assembly was dubbed a 'genome', a near complete gene set integrated into a chromosome-scale assembly using physical and genetic information. Sequence contigs that could not be assigned chromosomal positions in this way were binned by homology to low-coverage shotgun sequence of flow-sorted chromosome arms (Muñoz-Amatriaín et al. 2011). This method integrated 22% of the total assembled sequence by length, covering 91% of the genes, into the chromosome-scale assembly.

Bread wheat is a major global cereal grain, essential to human nutrition. The bread wheat genome is hexaploid, with a size estimated at approximately 17 Gbp, composed of three closely related and independently maintained genomes (the A, B and D genomes). This complex structure has resulted from two independent hybridization events. The first event brought together the diploid *T. urartu* (the A-genome donor) and an unknown *Aegilops* species, thought to be related to *Aegilops speltoides* (the B-genome donor), forming the allotetraploid *T. turgidum* around 0.5 million years ago (MYA). This species has produced both the emmer and durum wheat cultivars, the latter still being grown today for pasta. The second hybridization event brought together *T. turgidum* with *Ae. tauschii* (the D-genome donor) about 8,000 years ago in the Fertile Crescent.

Ensembl Plants contains version 1.0 of the chromosome survey sequence for *T. aestivum* cv. Chinese Spring, generated by the International Wheat Genome Sequencing Consortium (2014). The draft assembly of this complex genome was made tractable by using flow-sorted chromosome arms (Dolezel et al. 2007, Vrána et al. 2012). The assembly of gene-containing regions is reasonably good, with an N50 of 2.5 kb (**Table 2**).

Fig. 1 Visualizing the bread wheat genome through the Ensembl Genomes browser interface. The user can view many layers of genome annotation in a highly customizable way. Tracks shown include (A) gene models, (B) assemblies and interhomoeologous variations from Brenchley et al. (2012), (C) RNA-Seq data, (D) variations from the AXIOM array and (E) transcript assemblies from *T. turgidum*. Additional tracks are shown for *T. aestivum* ESTs and UniGenes (purple and green), alignment blocks to *O. sativa* and *B. distachyon* (pink), repeats (grey) and GC content.

The draft genome assemblies of *Ae. tauschii* (AL8/78) and *T. urartu* (G1812) are 4.23 and 4.66 Gbp, with N50 lengths of 58 and 64 kbp, respectively. They were both produced by the Chinese Academy of Agricultural Sciences using similar methodologies (Jia et al. 2013, Ling et al. 2013)

Data Import

Gene model annotations and gene names were imported from the relevant authority for each species (see references in **Table 1**). For specific genomes, additional sequence and variation data sets have been added, and are described below for the four Triticeae genomes in Ensembl Plants. Information about the analysis and visualization of the available data is described in the section 'Analysis and Visualization'.

Hordeum vulgare (barley)

A total of 79,379 gene models were described with the release of the barley genome (International Barley Genome Sequencing Consortium 2012). These models are classified as either high confidence (26,159 genes) or low confidence (53,220 genes), which are displayed in separate tracks in the browser. Only the high-confidence gene models were used for downstream analysis (see 'Analysis and Visualization', below). The gene

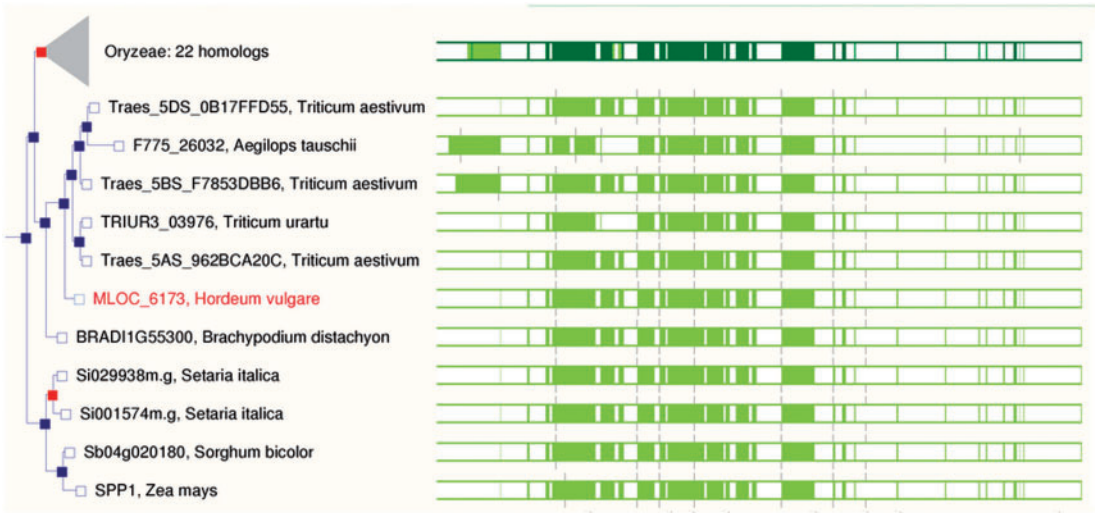

Fig. 2 Detailed view of a gene tree in Ensembl Plants. The tree shows the inferred evolutionary history of the sucrose-6F-phosphate phosphohydrolase family protein in *H. vulgare*. The gene tree (left) shows the expected phylogenetic relationship for the gene between the species shown. Note that the sequence identifier of the wheat genes includes the name of the chromosome arm to which it belongs, i.e. 5DS for the short arm of chromosome 5 in the D-genome. Red squares indicate inferred duplication events in the history of the gene, and shaded gray triangles indicate collapsed branches. A pictographic representation of the underlying multiple sequence alignment is included on the gene tree pages (right).

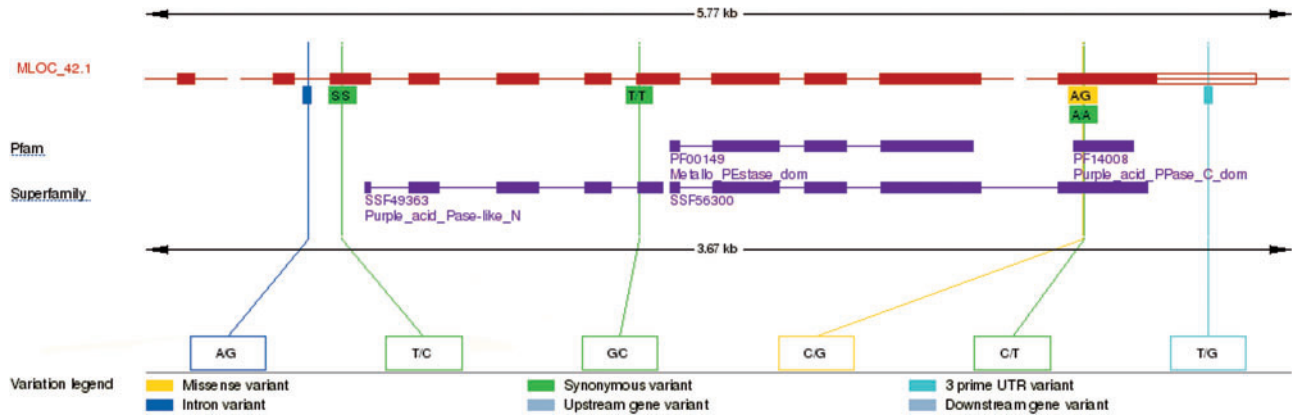

Fig. 3 The transcript variation image for the *H. vulgare* MLOC_42.1 protein-coding transcript in Ensembl Plants. The image gives an overview of all the variants within the transcript in the context of the functional domains assigned to the protein. Upper boxes highlight the amino acid change, where applicable, and lower boxes give the alleles. Variants are color coded according to their consequence type, missense, synonymous and positional. A full list of consequence types is given here: http://www.ensembl.org/info/genome/variation/predicted_data.html. The transcripts, features and variations can be clicked to explore more information about each object.

models for barley were made available through the MIPS barley genome database (http://mips.helmholtz-muenchen.de/plant/barley/).

Expressed sequences including expressed sequence tags (ESTs) from the HarvEST database (http://harvest.ucr.edu) and a set of non-redundant barley full-length cDNAs (Matsumoto et al. 2011) were aligned to the genome to demonstrate support for the gene models. Sequences from the Affymetrix GeneChip Barley Genome Array (http://www.affymetrix.com/catalog/131420/AFFY/Barley-Genome-Array) were also aligned, allowing users to search the genome by probe identifier and find the corresponding regions or transcripts in barley.

RNA sequencing (RNA-Seq) data were aligned from two studies: (i) SNP discovery in nine lines of cultivated barley (Study ERP001573) and (ii) RNA-Seq study of eight growth stages (Study ERP001600), both described in the reference publication (International Barley Genome Sequencing Consortium 2012).

Sequence variation was loaded from two sources: (i) variants derived from the whole-genome shotgun survey sequencing four cultivars, Barke, Bowman, Igri, Haruna Nijo and a wild barley, *H. spontaneum*; and (ii) variants derived from RNA-Seq from the embryonic tissues of nine spring barley varieties (Barke, Betzes, Bowman, Derkado, Intro, Optic, Quench, Sergeant and Tocada). Both approaches are described in the reference publication (International Barley Genome Sequencing Consortium 2012). **Fig. 3** shows an example view of barley variations in Ensembl Plants. See below for more information on analysis of variation data.

Triticum aestivum (bread wheat)

A total of 99,386 gene models were described with the release of the wheat chromosome survey sequence (International Wheat Genome Sequencing Consortium 2014). Their structure was computed by spliced-alignments of publically available wheat full-length cDNAs and the protein sequences of the related grass species, barley, Brachypodium, rice and Sorghum. A comprehensive RNA-Seq data set including five different tissues, root, leaf, spike, stem and grain, and different developmental stages was also used to identify wheat-specific genes and splice variants (**Fig. 1A**). The gene models for wheat were made available through the MIPS Wheat Genome Database (http://mips.helmholtz-muenchen.de/plant/wheat/).

A set of wheat genome assemblies (Brenchley et al. 2012) were aligned to the International Wheat Genome Sequencing Consortium assembly as well as to Brachypodium, barley and the wheat progenitor genomes. Homoeologous variants that were inferred between the three component wheat genomes in the same study are also displayed in Ensembl Plants in the context of the gene models of these five species (**Fig. 1B**).

RNA-Seq data were aligned from three studies: (i) Discovery of SNPs and genome-specific mutations by comparative analysis of transcriptomes of hexaploid wheat and its diploid ancestors (Study SRP002455; Akhunova et al. 2010); (ii) 454 sequencing of the *T. aestivum* cv. Chinese spring transcriptome (Study ERP001415; Brenchley et al. 2012); and (iii) *T. aestivum* transcriptome or gene expression (Study SRP004502) (**Fig. 1C**).

Variations for bread wheat were loaded from the CerealsDB (Wilkinson et al. 2012). A total of approximately 725,000 single nucleotide polymorphisms (SNPs) across approximately 250

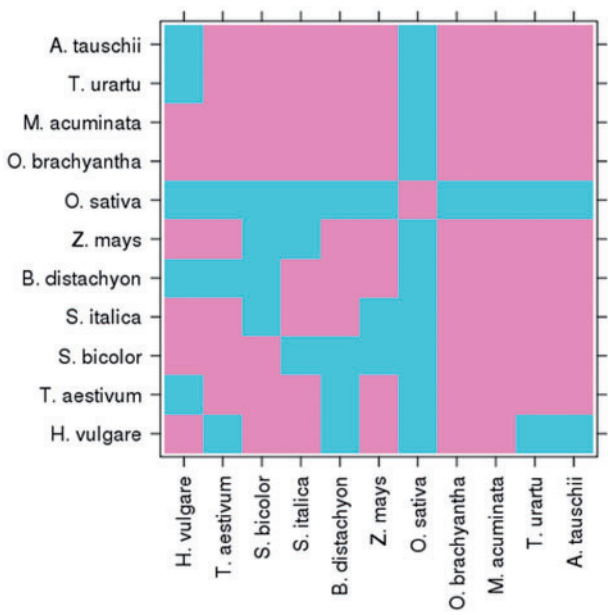

Fig. 4 The matrix of whole-genome alignments between pairs of monocot genomes in Ensembl Plants. Cyan indicates that an alignment exists for the pair. Only one representative rice is shown, *O. sativa*, although each of the 10 rice genomes was aligned against each other (not shown).

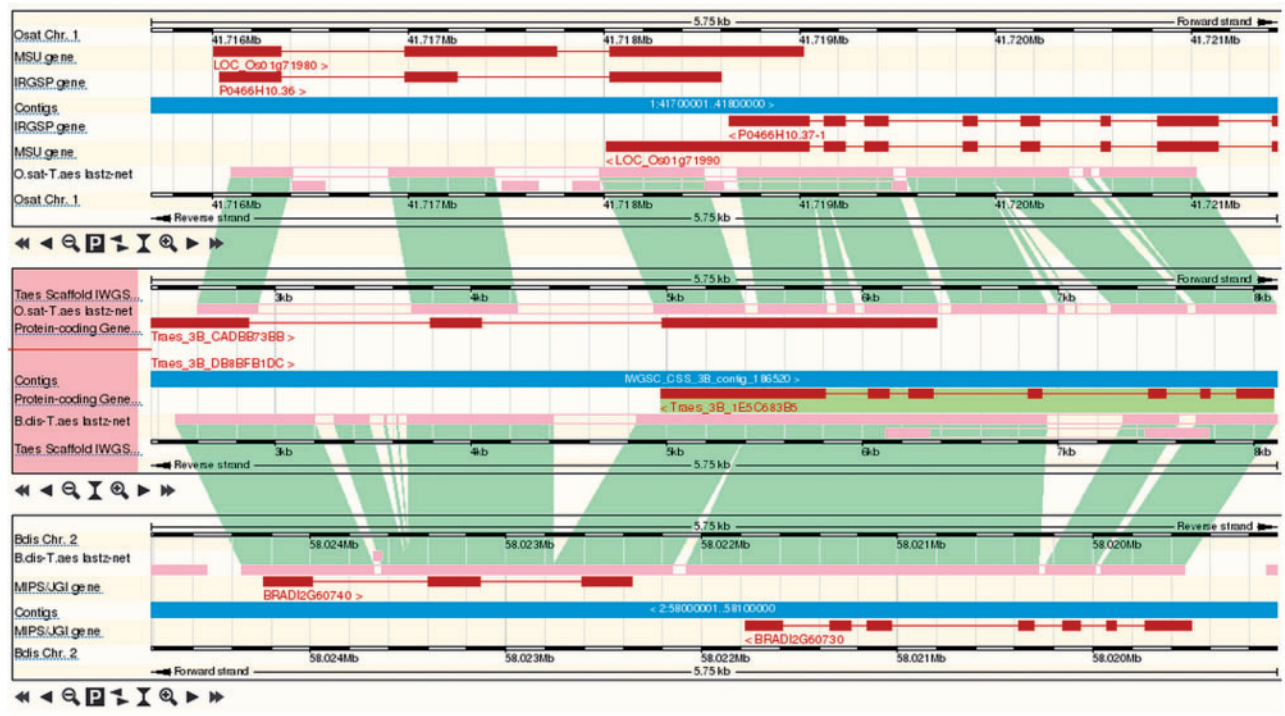

Fig. 5 View of the whole-genome alignment between wheat, rice and Brachypodium in Ensembl Plants. Pink bars and green blocks indicate aligned blocks between the rice and wheat (upper) and wheat and Brachypodium (lower) pairs of genomes. Transcripts are shown in red but the genomic features shown on each track are configurable.

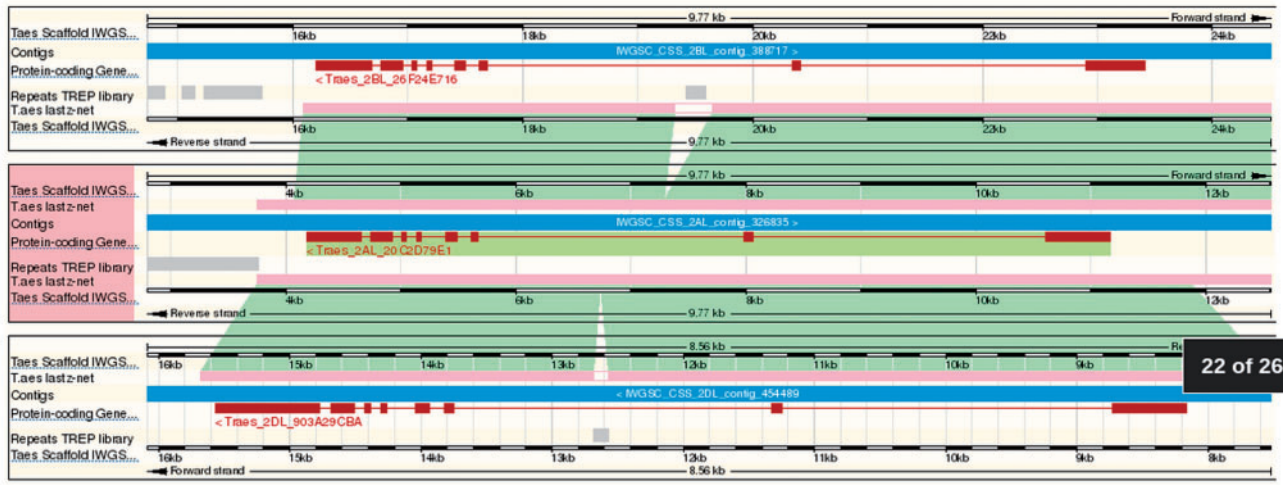

Fig. 6 Polyploid view of the whole-genome alignment within the bread wheat A, B and D component genomes. The image is defined as in **Fig. 5**. An additional feature track shows repeats annotated in all three genomes.

Table 1 Triticeae genomes in Ensembl Plants

Species (strain)	Description	Estimated genome size Mb	Assembly size Mb	Genes
Barley, *Hordeum vulgare* (cv. Morex)	Barley is an economically important crop and an important model of environmental diversity for development of wheat (International Barley Sequencing Consortium 2012).	5,100 (Doležel et al. 1998)	4,706	24,211
Bread wheat, *Triticum aestivum* (cv. Chinese spring)	An economically important food crop, accounting for >20% of global agricultural production (International Wheat Genome Sequencing Consortium 2014).	16,974 (Bennett and Smith 1991)	4,460	108,569
A-genome progenitor, *Triticum urartu* (G1812)	An einkorn wheat and the diploid progenitor of the bread wheat A-genome (Ling et al. 2013).	4,940 (Ling et al. 2013)	3,747	34,843
D-genome progenitor, *Aegilops tauschii* (AL8/78)	The diploid progenitor of the bread wheat D-genome (Jia et al. 2013).	4,360 (Jia et al. 2013)	3,314	33,849

Table 2 Some gene model statistics

Species	Contig N50 (kbp)	Average gene length	Average exon number	% complete genes	% InterPro coverage
O. sativa		3,090	2.5	77	68
B. distachyon		3,582	5.6	99	86
H. vulgare	0.9/3.2	2,812	5.4	76	84
T. urartu	3.4/5.8	3,208	4.7	78	78
Ae. tauschii	4.5/6.2	2,935	4.9	100	77
T. aestivum	2.4/6.3	2,197	3.8	56	84

Contig N50 reported twice, once for the complete assembly/and once for just the gene-containing contigs.
Complete genes are defined as those starting with a methionine, ending in a stop codon.
InterPro coverage consists only of structural protein domains and functional motifs, excluding low complexity, coiled-coil, transmembrane and signal motifs.

varieties were loaded. These SNP loci are associated with markers from three genotyping platforms: the Axiom 820K SNP Array, the iSelect 80K Array (Wang et al. 2014) and the KASP probe set (Allen et al. 2011). In addition to these intervarietal SNPs, work is ongoing to generate and report interhomoeologous variants (**Fig. 1D**).

Tritcum urartu and *Aegilops tauschii* (the bread wheat A and D progenitor genomes)

A total of 34,879 and 34,498 protein-coding genes were reported for *T. urartu* and *Ae. tauschii*, respectively. They were predicted using FGENESH (Salamov and Solovyev 2000) and

Table 3 A list of the standard computational analyses that are routinely run over all genomes in Ensembl Plants

Pipeline name	Summary
Repeat feature annotation	Three repeat annotation tools are run, RepeatMasker, Dust and TRF. RepeatMasker was run with repeat libraries from Repbase as well as Triticeae specific repeats from TREP. http://ensemblgenomes.org/info/data/repeat_features
Non-coding RNA annotation	tRNAs and rRNAs are predicted using tRNAScan-SE (Lowe and Eddy 1997) and RNAmmer (Lagesen et al. 2007), respectively. Other ncRNA types are predicted by alignment to Rfam models (Griffiths-Jones 2005). http://ensemblgenomes.org/info/data/ncrna
Feature density calculation	Feature density is calculated by chunking the genome into bins, and counting features of each type in each bin.
Annotation of external cross-references	Database cross references are loaded from a predefined set of sources for each species, using either direct mappings or by sequence alignment. http://ensemblgenomes.org/info/data/cross_references
Ontology annotation	In addition to database cross references, ontology annotations are imported from external sources (Ashburner et al. 2000, Cooper et al. 2013). Terms are additionally calculated using a standard pipeline based on domain annotation (Jones et al. 2014) and are projected between orthologs defined by gene tree analyis. http://ensemblgenomes.org/info/data/cross_references
Protein feature annotation	Translations are run through InterProScan (Jones et al. 2014) to provide protein domain feature annotations. http://ensemblgenomes.org/info/data/protein_features
Gene trees	The peptide comparative genomics (Compara) pipeline (Vilella et al. 2009) computes feature-rich gene trees for every protein in Ensembl Plants. http://ensemblgenomes.org/info/data/peptide_compara
Whole-genome alignment	Whole-genome alignments are provided for closely related pairs of species based on LastZ or translated BLAT results. Where appropriate, K_a/K_s and synteny calculations are included. http://ensemblgenomes.org/info/data/whole_genome_alignment
Short read alignment	Short reads are automatically downloaded from the SRA by study accession and are aligned to a given reference by using BWA (Li and Durbin 2009), GSNAP (Wu and Nacu 2010) or STAR (Dobin et al. 2013), depending on the characteristics of the library. Alignments are stored in BAM or WIG format, and may be viewed as coverage or density tracks in the browser. http://ensemblgenomes.org/info/data/short_read_mapping
Variation coding consequences	For those species with data for known variations, the coding consequences of those variations are computed for each protein-coding transcript (McLaren et al. 2010). http://plants.ensembl.org/info/docs/tools/vep/index.html

GeneID (Guigó et al. 1992) with supplemental evidence from RNA-Seq and EST sequences (Jia et al. 2013, Ling et al. 2013). In addition, approximately 200,000 bread wheat UniGene cluster sequences were aligned to both genomes using Exonerate (Slater and Birney 2005). UniGene cluster sequences are based on reads from cDNA and EST libraries across a variety of samples (Wheeler et al. 2003). Similar sequences are clustered into transcripts and, in this case, are filtered by species (*T. aestivum*). Similarly, all publicly available bread wheat ESTs, retrieved using the European Nucleotide Archive (Leinonen et al. 2011) advanced search, were aligned using STAR (Dobin et al. 2013).

For *Ae. tauschii*, RNA-Seq data were aligned from a single study: RNA-Seq from seedling leaves of *Ae. tauschii* (Study DRP000562; Iehisa et al. 2012).

Transcriptomes

In addition to the four Triticeae genomes, the transcriptome assembly of *Triticum turgidum* (durum wheat) is presented by alignment to *T. aestivum* and the transcriptome assemblies of two *Triticum monococcum* (einkorn wheat) subspecies are presented by alignment to *T. aestivum*, *T. urartu* and *H. vulgare* (**Fig. 1E**). These resources contain between 118,000 and 140,000 transcripts each. The alignments to the selected reference genomes allows comparative analysis to be performed between the different resources, including 'lift-over' of features such as SNPs from one species to another, described in Fox et al. (2014).

Analysis and Visualization

Every genome hosted by Ensembl Plants is subject to several automatic computational analyses, summarized in **Table 3**. Some of the key analysis and their resulting visualizations are described in more detail below.

Whole-genome alignments

A total of 55 pairwise whole-genome alignments are provided for the 20 monocot genomes in Ensembl Plants (**Fig. 4**). Pairs include bread wheat against barley, *T. urartu* against *Ae. tauschii*, and *Sorghum bicolor* against *Zea mays* and barley. Each genome was aligned to the *Oryza sativa* (Japonica) genome, allowing any pair of genomes to be indirectly compared via this reference. Additional comparisons include an all-against-all comparison of the 10 rice genomes, produced in collaboration with Gramene (Monaco et al. 2014).

In the first step of the whole-genome alignment pipeline, two types of pairwise alignment may be used: either LastZ (Harris, 2007), for closely related species, or translated BLAT (Kent, 2002), typically for more distantly related species.

After the initial alignment step, non-overlapping, collinear 'chains' of alignment blocks are identified, and the final step 'nets' together compatible chains to find the best overall alignment (Kent et al. 2003). When sequence similarity between the pair is sufficiently high, the ratio of the number of non-synonymous substitutions per non-synonymous site (dN) to the number of synonymous substitutions per synonymous site (dS), which can be used as an indicator of selective pressure acting on a protein-coding gene (dN/dS), and synteny calculations are included and may be visualized in the genome browser.

Whole-genome alignments are used to support the parallel visualization of aligned genomic regions across multiple related species in the browser in the 'Region Comparison View' (**Fig. 5**), allowing the inspection of conserved features and differences such as gene structure, copy number, polymorphism and repeat content between the genomes of multiple species.

Another potential use of whole-genome alignments is the creation of functional 'projection assemblies' between a genome with a chromosome-scale assembly, such as Brachypodum or barley, to one currently without, such as wheat. This task may be performed using BioMart in several steps. For example, two gene-based wheat markers may span a quantitative trait locus (QTL) of interest, but would probably not be located on the same contiguous assembly in the fragmented draft bread wheat assembly. However, one can retrieve the orthologs of these genes on the barley assembly where they are likely to belong to the same chromosomal region. In a second step, all the wheat orthologs of the barley genes in the region can be retrieved.

Polyploidy

Building on the region comparison view for whole-genome alignments, a recently developed 'Polyploid View' allows users to browse homoeologous genomic features on the wheat A, B and D component genomes in parallel (**Fig. 6**). Alignments between contigs containing homoeologous gene family members, identified using gene trees (described below), can be directly visualized from the 'Homoeologues' page for each gene, and can be visualized with a single click. These alignment data are generated by comparing the three bread wheat genomes with each other using the same protocol as that used for the interspecies alignments. An additional filtering step retains only those alignments containing genes with inferred 'orthology' between the component genomes (defined as homoeologs, as described below). By using this definition, paralogous alignments between gene families are not shown in the polyploid view.

Gene trees

The Ensembl Gene Tree pipeline is used to calculate evolutionary relationships among members of protein families (Vilella et al. 2009). Clusters of similar sequences are first identified using BLAST+ (Camacho et al. 2009), then each cluster of proteins is aligned using M-Coffee (Wallace et al. 2006) or, when the cluster is very large, Mafft (Katoh et al. 2002). Finally, TreeBeST (Vilella et al. 2009) is used to produce a gene tree from each multiple alignment by reconciling the relationship between the sequences with the known evolutionary history to call gene duplication events. This final step allows the identification of orthologs, paralogs and, in the case of polyploid genomes, homoeologs.

TreeBeST merges five input trees into one tree by minimizing the number of duplications and losses into one consensus tree. This allows TreeBeST to take advantage of the fact that DNA-based trees often are more accurate for closely related parts of trees and protein-based trees for distant relationships, and that a group of algorithms may outperform others under certain scenarios. The algorithm simultaneously merges the five input trees into a consensus tree. The consensus topology contains clades not found in any of the input trees, where the clades chosen are those that minimize the number of duplications and losses inferred, and have the highest bootstrap support.

The resulting gene tree is an evolutionary history of a gene family, including identification of candidate gene duplication and speciation events, derived from the multiple sequence alignments. From the gene tree, one can identify true orthologs, paralogs and, in the case of polyploid genomes, homoeologs. Part of an example gene tree is shown in **Fig. 2**, showing the inferred evolutionary history of a protein in the sucrose-6F-phosphate phosphohydrolase family, including speciation and duplication events.

As a relatively large set of closely related genomes, the Poaceae (true grasses, of which the Triticeae are a subset) are particularly interesting. Using the gene tree analysis, we have placed 690,172 cereal genes into 39,216 groups of implied orthology. In spite of the provisional nature of many of these genome assemblies, many of these orthologous groups are represented in every genome. A total of 7,203 groups (containing 260,004 genes) cover all 21 Poaceae genomes (counting the bread wheat A, B and D component genomes separately); 18,433 orthologous groups cover between two and 21 genomes with a single representative from each genome in the group; and 954 groups contain a single representative from all 21 genomes.

One of the benefits of extensive and accurate prediction of orthologs across plant species is the ability to project functional annotation between pairs of orthologous genes on the assumption that orthologs generally retain function between species (Altenhoff et al. 2012). Using this methodology we have projected manually curated gene ontology (GO) terms from *O. sativa* to the other monocots. Projected terms are tagged as 'inferred from electronic annotation' (IEA) to prevent confusion with curated GO terms resulting from direct experimental evidence. The bulk of GO annotations for most species before projection are IEA assignments that come from Interpro2Go (see **Table 2**) that tend to be functionally broad. In contrast, projected terms can provide far more detailed annotation.

Variation

The Ensembl Plants variation module is able to store variant loci and their known alleles, including SNPs, indels and

structural variations; the functional consequence of known variants on protein-coding genes; and individual genotypes, population frequencies, linkage and statistical associations with phenotypes. In the case of the polyploid bread wheat genome, heterozygosity, intervarietal variants and interhomoeologous variants are stored and visualized distinctly. A variety of views allow users to access these data (e.g. **Fig. 3**) and variant-centric warehouses are produced using BioMart. In addition, the Variant Effect Predictor allows users to upload their own data and see the functional consequence of self-reported variants on protein-coding genes (McLaren et al. 2010).

Future Directions

The rapid pace of progress in the field of cereal genomics is driving the continued development of Ensembl Plants. Triticeae resources are prioritized within the resource, and we aim to include new data sets rapidly after publication and data release. At the same time, the complex nature of these genomes is necessitating ongoing improvements to our analysis pipelines and user interface.

Specific developments that are planned within the next few months include the release of improved genomic assemblies for barley and wheat. Although these assemblies will not be contiguous and ungapped, the use of additional genetic data will allow the approximate positioning and orientation of a larger number of genes within a chromosome-scale framework.

We expect to release an update to the barley genome assembly in release 25, due in January 2015, using a new marker set derived from population sequencing (POPSEQ) to anchor sequence contigs to chromosomal locations (Mascher et al. 2013). The new assembly will anchor an additional 346 Mb of sequence (411,526 contigs), containing 995 genes (5% of the total).

Similarly, we expect progress in the development of the wheat genome assembly towards full chromosome assemblies that makes use of both the recently released sequence of the 3B chromosome (Choulet et al. 2014) and integration of POPSEQ data across the whole genome (Poland et al. 2012, International Wheat Genome Sequencing Consortium unpublished). Fuller assemblies will alleviate the computational burden of whole-genome alignment, which is problematic when genomes are highly fragmented, allowing for the maintenance of a wider range of whole-genome comparisons.

In addition to expanded variation data for bread wheat, we expect that the whole-genome alignments between the A, B and D component genomes will be used to generate an extensive set of intervarietal variations, and will be added to the existing variation data. Similar interspecies analysis based on the *T. monococcum* transcriptome will provide yet another source of wheat variation data.

In the longer term, we plan to extend the range and scope of RNA sequence alignments to the plant genomes hosted in Ensembl Plants by developing automatic methods to discover the relevant entries in the European Nucleotide Archive (Leinonen et al. 2011) based on their descriptive metadata. Matching data sets will be aligned automatically, and a new configuration interface will allow users to select studies to view against the relevant genomes based on matching search criteria against submission metadata. Work is also ongoing to integrate data and visualization tools from the ArrayExpress (Rustici et al. 2013) and Atlas (Petryszak et al. 2014) resources into the browser, to allow expression data between tissues, time-series or species to be viewed in a consistent way.

Funding

This work was supported by Biotechnology and Biological Sciences Research Council [BB/I008357/1, BB/J003743/1]; the 7th Framework Programme of the European Union [283496].

Acknowledgments

Ensembl Plants is funded as part of the transPLANT project within the 7th Framework Programme of the European Union, contract number 283496. Databases are constructed in a direct collaboration with the Gramene resource, funded by the US National Science Foundation award #1127112. D.M.B. wishes to thank Cristobal Uauy for the 'projection assemblies' use case.

Disclosures

The authors have no conflicts of interest to declare.

References

Akhunova, A.R., Matniyazov, R.T., Liang, H. and Akhunov, E.D. (2010) Homoeolog-specific transcriptional bias in allopolyploid wheat. *BMC Genomics* 11: 505.

Allen, A.M., Barker, G.L.A., Berry, S.T., Coghill, J.A., Gwilliam, R., Kirby, S., Robinson, P. et al. (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (Triticum aestivum L.). *Plant Biotechnol. J.* 9: 1086–1099.

Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M. and Dessimoz, C. (2012) Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput. Biol.* 8: e1002514.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M. et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25: 25–29.

Bennett, M.D. and Smith, J.B. (1991) Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. B: Biol. Sci* 334: 309–45.

Brenchley, R., Spannagl, M., Pfeifer, M., Barker, G.L.A., D'Amore, R., Allen, A.M. et al. (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491): 705–710.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10: 421.

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J. et al. (2014) Structural and functional partitioning of bread wheat chromosome 3B. *Science* 345: 1249721.

Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B. et al. (2013) The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant Cell Physiol* 54: e1.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S et al. (2013) STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* 29: 15–21.

Doležel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M.A., Nardi, L. et al. (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.* 82: 17–26.

Dolezel, J., Kubaláková, M., Paux, E., Bartos, J. and Feuillet, C. (2007) Chromosome-based genomics in the cereals. *Chromosome Res.* 15: 51–66.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S. et al. (2014) Ensembl 2014. *Nucleic Acids Res.* 42: D749–D755.

Fox, S.E., Geniza, M., Hanumappa, M., Naithani, S., Sullivan, C., Preece, J et al. (2014) De novo transcriptome assembly and analyses of gene expression during photomorphogenesis in diploid wheat Triticum monococcum. *PLoS One* 9: e96855.

Goff, S.A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296: 92–100.

Griffiths-Jones, S. (2005) Annotating non-coding RNAs with Rfam. *Curr. Protoc. Bioinformatics* Chapter 12: Unit 12.5.

Guigó, R., Knudsen, S., Drake, N. and Smith, T. (1992) Prediction of gene structure. *J. Mol. Biol.* 226: 141–157.

Harlan, J.R. and Zohary, D. (1966) Distribution of wild wheats and barley. *Science* 153: 1074–1080.

Harris, R.S. (2007) Improved Pairwise Alignment of Genomic DNA ProQuest.

Iehisa, J.C.M., Shimizu, A., Sato, K., Nasuda, S. and Takumi, S. (2012) Discovery of high-confidence single nucleotide polymorphisms from large-scale de novo analysis of leaf transcripts of Aegilops tauschii, a wild wheat progenitor. *DNA Res.* 19: 487–497.

International Barley Genome Sequencing Consortium, Mayer, K.F.X., Waugh, R., Brown, J.W.S., Schulman, A., Langridge, P., Platzer, M. et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 2012 491: 711–716.

International Wheat Genome Sequencing Consortium (IWGSC). (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science* 345: 1251788.

Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W. et al. (2013) Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496: 91–95.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C et al. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30: 1236–1240.

Kasprzyk, A. (2011) BioMart: driving a paradigm change in biological data management. *Database (Oxford)* 2011: bar049.

Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30: 3059–3066.

Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.* 12: 656–664.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. (2003) Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl Acad. Sci. USA* 100: 11484–11489.

Kersey, P.J., Allen, J.E., Christensen, M., Davis, P., Falin, L.J., Grabmueller, C. et al. (2014) Ensembl Genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res.* 42: D546–D552.

Lagesen, K., Hallin, P., Einar, A.R., Staerfeldt, H.-H., Rognes, T. and Ussery, D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* 35: 3100–3108.

Lai, K., Berkman, P.J., Tadeusz Lorenc, M., Duran, C., Smits, L., Manoli, S. et al. (2012) WheatGenome.info: an integrated database and portal for wheat genome information. *Plant Cell Physiol.* 53: e2.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y. et al. (2011) The European Nucleotide Archive. *Nucleic Acids Res.* 39: D28–D31.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C. et al. (2013) Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature* 496: 87–90.

Lowe, T.M. and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25: 955–964.

Mascher, M., Muehlbauer, G.J., Rokhsar, D.S., Chapman, J., Schmutz, J., Barry, K. et al. (2013) Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ). *Plant J.* 76: 718–727.

Matsumoto, T., Tanaka, T., Sakai, H., Amano, N., Kanamori, H., Kurita, K. et al. (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* 156: 20–28.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. and Cunningham, F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor. *Bioinformatics* 26: 2069–2070.

Monaco, M.K., Stein, J., Naithani, S., Wei, S., Dharmawardhana, P., Kumari, S. et al. (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.* 42: D1193–D1199.

Muñoz-Amatriaín, M., Moscou, M.J., Bhat, P.R., Svensson, J.T., Bartoš, J., Suchánková, P. et al. (2011) An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome J.* 4: 238.

Obayashi, T., Okamura, Y., Ito, S., Tadaka, S., Aoki, Y., Shirota, M. and Kinoshita, K. (2014) ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants. *Plant Cell Physiol.* 55: e6.

Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N.A., Gonzalez-Porta, M, Hastings, E. et al. (2014) Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res.* 42: D926–D932.

Poland, J.A., Brown, PJ., Sorrells, M.E. and Jannink, J.-L. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7: e32253.

Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I. et al. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.* 41: D987–D990.

Salamov, A.A. and Solovyev, V.V. (2000) Ab initio gene finding in *Drosophila* genomic DNA. *Genome Res.* 10: 516–522.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, JC., Wei, F., Pasternak, S. et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326: 1112–1115.

Slater, G.St.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6: 31.

UniProt Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 42: D191–D198.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.h* 19: 327–335.

Vrána, J., Simková, H., Kubaláková, M., Číhalíková, J. and Doležel, J. (2012) Flow cytometric chromosome sorting in plants: the next generation. *Methods* 57: 331–337.

Wallace, I.M., O'Sullivan, O., Higgins, D.G. and Notredame, C. (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.h* 34: 1692–1699.

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B.E. et al. (2014) Characterization of polyploid wheat genomic diversity using a high-density 90,000 single nucleotide polymorphism array. *Plant Biotechnol. J.* 12: 787–796.

Wheeler, D.L., Church, D.M., . Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U. et al. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28–33.

Wilkinson, P.A., Winfield, M.O., Barker, G.L.A., Allen, A.M., Burridge, A., Coghill, J.A. et al. (2012) CerealsDB 2.0: an integrated resource for plant breeders and scientists. *BMC Bioinformatics* 13: 219.

Wu, T.D. and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26: 873–881.

Yonemaru, J., Ebana, K. and Yano, M. (2014) HapRice, an SNP haplotype database and a web tool for rice. *Plant Cell Physioly* 5: e9.

Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B. et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* 296: 79–92.